



Abt Associates Inc.

# *memorandum*

**Date** February 2, 2005

**To** William E. Doyle, Jr.

**From** David C. Hoaglin

**Subject** Issues Arising from Review of Draft Report

We have investigated the main issues raised in the conference call on December 15, 2004, concerning the draft report "An Evaluation of OFCCP's Equal Opportunity Survey." This memo summarizes our findings.

## **Relation between Outcomes on the Survey and Final Dispositions from the Review Process**

Several questions focused on how the 6,400 establishments in the review subsample (whose dispositions are summarized in Table 3.2 of the report) and the 3,048 establishments whose surveys had status "OK" (Table 3.3) compare with the 10,018 establishments in the survey sample (Table 3.1). As a first step toward those comparisons, Table A separates the 10,018 establishments in the survey sample into the 6,400 that were in the review subsample and the 3,618 that were not in the review subsample. (Table A uses the same categories of final dispositions on the survey as Table 3.1.) The main difference between the two parts of the survey sample is that a substantially higher percentage of the establishments in the review subsample had surveys with status "OK" (48% versus 33%) and a correspondingly lower percentage fell into the category "Other surveys with data" (the sum of the percentages in these two categories is nearly the same in the two parts). The most likely explanation for this difference is that the data-receipt contractor made a special effort to resolve questions about the data submitted for establishments in the review subsample (usually by recontacting the respondent), and thus to move those surveys into the "OK" category.

More information on the relation between survey outcome and review outcome comes from cross-tabulating the 6,400 establishments in the review subsample by the two sets of categories. The row categories in Table B come from Table 3.1, and the column categories come from Table 3.2. Thus, the column totals in Table B reproduce the counts in Table 3.2, and the first row of Table B reproduces Table 3.3. Not surprisingly, the vast majority of establishments in the last three rows of Table B fall in the last column (Review never opened).

Of particular interest were the 22 (= 89 – 67) establishments in which the review found systemic discrimination but the 2002 EO Survey did not have status "OK." Also of interest were the 442 (= 2,601 – 2,159) in which the review found no systemic discrimination and the survey did not have status "OK." Most of these (20 of 22 and 405 of 442) belong to the second row of Table B (Other surveys with data). As we report further below, we were able to rerun a substantial part of our analysis after including the

data from 16 of the 20 and 299 of the 405. The remaining 4 and 106 had a final status code of ECRG (Edit Condition Report Generated) on the survey; we do not regard their data as trustworthy.

**Table A. Final Dispositions on the 2002 EO Survey, by Whether the Establishment Was in the Review Subsample**

	<b>In Review Subsample</b>		<b>Not in Review Subsample</b>	
<b>Final Disposition</b>	<i>n</i>	%	<i>n</i>	%
Status “OK”	3,048	47.6	1,206	33.3
Other surveys with data	675	10.5	847	23.4
Nonrespondents	673	10.5	331	9.1
Asserted no jurisdiction	1,674	26.2	1,064	29.4
Out of business	330	5.2	170	4.7
Total	6,400	100.0	3,618	100.0

**Table B.**

<b>Table of SVY_DISP by REV_DISP</b>						
<b>SVY_DISP (2002 EO Survey Disposition)</b>	<b>REV_DISP (Review Process Disposition)</b>					
<b>Frequency Percent Row Pct Col Pct</b>	<b>Systemic discrimination</b>	<b>No systemic discrimination</b>	<b>Review started but not completed</b>	<b>Not reviewable</b>	<b>Review never opened</b>	<b>Total</b>
<b>Status OK</b>	67 1.05 2.20 75.28	2159 33.73 70.83 83.01	9 0.14 0.30 40.91	378 5.91 12.40 63.85	435 6.80 14.27 14.05	3048 47.63
<b>Other surveys with data</b>	20 0.31 2.96 22.47	405 6.33 60.00 15.57	1 0.02 0.15 4.55	121 1.89 17.93 20.44	128 2.00 18.96 4.13	675 10.55
<b>Nonrespondents</b>	1 0.02 0.15 1.12	10 0.16 1.49 0.38	2 0.03 0.30 9.09	15 0.23 2.23 2.53	645 10.08 95.84 20.83	673 10.52
<b>Asserted no jurisdiction</b>	0 0.00 0.00 0.00	24 0.38 1.43 0.92	9 0.14 0.54 40.91	70 1.09 4.18 11.82	1571 24.55 93.85 50.74	1674 26.16
<b>Out of business</b>	1 0.02 0.30 1.12	3 0.05 0.91 0.12	1 0.02 0.30 4.55	8 0.13 2.42 1.35	317 4.95 96.06 10.24	330 5.16
<b>Total</b>	89 1.39	2601 40.64	22 0.34	592 9.25	3096 48.38	6400 100.00

## Impact of Including Additional Data

As mentioned above, we included the 315 (= 16 + 299) establishments whose surveys had data but not status “OK” and then repeated a substantial part of the model-building process. The result, in brief, was that we emerged with the same four predictor variables. The coefficients were somewhat different, but not greatly so. The qualitative interpretation is pretty much the same.

More specifically, we derived values of the predictor variables for the 315 additional establishments. We then compared those 315 establishments against the earlier 2,226 on the 22 predictor variables that we had used as the basis for building the stepwise logistic model. (These comparisons separated the establishments with findings of systemic discrimination from those with findings of no systemic discrimination.) On three of the predictors derived from Part B, the additional establishments included an outlier (relative to the data from the original establishments); but otherwise the additional data were generally similar to the original data, often showing reduced variation, as one would expect with the smaller sample sizes.

We then returned to the longer list of predictor variables that, in the numerical and graphical single-variable analyses, had shown some relation to SD. We again considered each of these variables separately in a single-variable logistic regression model, now using the augmented data (2,541 = 2,226 + 315 establishments). This step identified 18 predictor variables whose p-value in their single-variable logistic regression was less than 0.25: five variables on the earlier list of 22 (Table 5.1) were not on this list, and one variable on this list was not on the earlier list.

When we reran the stepwise logistic regression, starting with the new list of 18 predictor variables, the resulting model contained three of the four predictors in the earlier model (Table 5.2). The fourth variable was `CompFemMale_TenureRatio`. This model was based on data from 1,700 establishments (i.e., it used data from 210 of the 315 additional establishments).

For comparison, we then fitted the earlier model (four predictors) to the augmented data. Any establishments that had no missing values on the four predictors, but missing values on one or more of the other 14 predictors, were now included --- with one exception. One of the additional establishments with  $SD = N$  had `MinWhite_TenureRatio` = 43.7, an extreme outlier; we set that observation aside (along with the six observations set aside earlier). The number of observations used was then 2,153, of which 78 had  $SD = Y$  (i.e., 15 of the 16 additional establishments with  $SD = Y$  contributed data). All four predictors had p-values below .05 (`CompFemMale_TenureRatio` had  $p = .0298$ ). Table C shows the coefficients, standard errors, and p-values. To facilitate comparisons, we also include the corresponding table of the report (Table 5.4). For each variable the coefficients for the augmented data are quite similar to those for the original data; each difference is smaller than the standard error in Table C. (We expect those standard errors to be smaller than the ones in Table 5.4 because the number of observations is larger.)

**Table C. Final Logistic Regression Model, Fitted to the Augmented Data (2,153 observations were used)**

Variable	Coefficient	Standard Error	P-Value
Intercept	-4.908	1.112	<.0001
Indicator_GT200	1.150	0.301	.0001
MinWhite_TenureRatio	-1.242	0.434	.0042
FemMale_Diffi3	-9.601	3.183	.0026
CompFemMale_TenureRatio	2.479	1.141	.0298

**Table 5.4 Refitted Model with Four Predictor Variables after Setting Aside 6 of the 2,226 Observations (1,888 observations were used)**

Variable	Coefficient	Standard Error	P-value
Intercept	− 5.318	1.312	<.0001
Indicator_GT200	1.193	0.339	.0004
MinWhite_TenureRatio	− 1.492	0.506	.0032
FemMale_Diffi3	− 10.685	3.726	.0041
CompFemMale_TenureRatio	3.040	1.345	.0239

### Missing Values on Compensation Predictors

Another concern expressed in the conference call was that the total absence from the models of any predictors derived from compensation might be the result of missing values on those predictors. The explanation, however, does not seem to lie in the numbers of missing values. For the original data, Table D lists the numbers of missing values among the 67 establishments with  $SD = Y$  and the 2,159 establishments with  $SD = N$ . In this respect the compensation variables did not stand out from other predictors. The tenure variables, also derived from data from Part C, had exactly the same counts of missing values for FemMale and MinWhite among the establishments with  $SD = Y$  and slightly smaller counts (168 and 230, respectively) among the establishments with  $SD = N$ . The 93 predictor variables derived from data in Part B showed substantially more variation. For establishments with  $SD = Y$  the number of missing values ranged from 0 to 13 and exceeded 2 on 40 predictors. For establishments with  $SD = N$  the number of missing values ranged from 0 to 57 on 53 predictors and from 230 to 446 on the remaining 40 predictors.

**Table D. Number of Establishments with Missing Values on the Predictors Derived from the Compensation Data**

Predictor	$SD=Y$	$SD=N$
FemMale_AAWRatio	1	187
FemMale_AAWRatioMinimum	1	187
MinWhite_AAWRatio	2	252
MinWhite_AAWRatioMinimum	2	252
CompFemMale_AAWRatio	1	189
CompFemMale_AAWRatioMinimum	1	189
CompMinWhite_AAWRatio	2	252
CompMinWhite_AAWRatioMinimum	2	252
Total number of establishments	67	2,159

Perhaps the questionable compensation data reported by some establishments reduce the predictive potential of the compensation variables. Alternatively, presence of systemic discrimination may not be associated with these measures of compensation.

### Applying Sampling Weights to the Completed Reviews

All analyses reported in the draft report were unweighted, but the main sample was based on 276 strata, and the review subsample was drawn by using another set of three strata. The objective of the main sample was to give larger establishments a higher probability of selection, as this would increase the

chance of having establishments with SD in the review subsample. Thus, a number of questions in the conference call had to do with the sample design and the possible impact of applying sampling weights to the data. We have addressed these questions by developing weights for the establishments that were in scope and had a completed compliance review. These weights incorporated a base weight from the selection of the main sample, a factor that reflected the sampling fraction in the selection of the review subsample, and an adjustment for nonresponse on the compliance review.

As discussed in Chapter 2 of the draft report, we obtained the main sample by drawing a simple random sample from each of the 276 final strata (after allocating sample among strata in proportion to the total number of employees reported). Appendix A of the report gives, for each stratum, the number of establishments in the stratum and the number of establishments in the sample. The corresponding base sampling weight equals the ratio of those two quantities (i.e., it equals the reciprocal of the sampling fraction for the stratum). Not surprisingly, the base weights vary strongly among the four categories of SIZE that were used in defining the strata. For all strata where the number of employees was 500 or more, the base weight equals 1.0. At the other end of the range, for strata where the number of employees was less than 150, it ranged from 4.33 to 5.52. In between, the largest value for strata where the number of employees was 300 – 499 was 1.25; and the values for strata where the number of employees was 150 – 299 ranged from 1.84 to 2.18.

For the three strata used in selecting the review subsample, the reciprocal of the sampling fraction is equal to the size of the selection interval in the systematic random sampling (for this purpose we treat the third part of the sample as if it had been selected all at once, instead of in three draws): 11.647 for the first part, 6.0 for the second part, and 1.4251 for the third part.

In adjusting for nonresponse on the compliance review (i.e., for establishments that were in scope but did not have a completed review), it was clear that many of the 276 strata would have too few completes to support a reasonably stable adjustment. Since 2,690 establishments had a completed review, the general dimensions of the problem are clear; the rightmost column of Appendix A in the report gives the details. As a reasonable compromise we used a set of 47 adjustment cells, based on a two-variable margin of the three-variable array of strata (12 categories of industry by 4 categories of size, less 1 because we combined a cell that contained only 5 completes with an adjacent cell). Each of the 47 cells contained at least 10 completes.

In preparation for calculating the nonresponse-adjustment factors, we classified as out-of-scope those establishments that had asserted no jurisdiction or were out of business or whose survey had been returned by the post office. We then assigned each in-scope establishment in the review subsample a preliminary sampling weight, equal to the product of the base weight for its stratum of the main sample and the factor for its part of the subsample. The provisional value of the adjustment factor for a cell was equal to the ratio of the total preliminary weight in the cell to the corresponding sum of the preliminary weight of the establishments with completed reviews. In four of the 47 cells the provisional adjustment factor was noticeably higher than in the other cells (those four ranged from 3.56 to 6.04, whereas the other 43 ranged from 1.25 to 2.87). Thus, to avoid potential problems with inflated variances and increased influence in logistic regressions, we truncated the four high factors to 3.0.

Even with this truncation the adjusted weights had an uncomfortably wide range, from 1.79 to 96.4. Not surprisingly, the main source of this variation was the preliminary weights (which ranged from 1.43 to 55.53) --- more specifically, the subsampling factors (as discussed above). Because there are only three subsampling factors, modifying them would have broad implications. Instead, we modified the base

weights by imposing an upper limit of 3.0 (this affected only establishments in the smallest of the four size categories, which ordinarily have low priority for compliance reviews). The resulting preliminary weights ranged from 1.43 to 34.94, and the corresponding “modified adjusted weights” ranged from 1.79 to 60.53.

We used these modified adjusted weights to refit the four-variable logistic regression model (Table C) to the augmented data, producing the model in Table E. The coefficients and standard errors in Table E differ somewhat from those in Table C, but not dramatically. A further comparison examined the relation between the predicted probabilities (of SD) from these two models. When the predicted probabilities from the weighted model (Table E) were plotted against those from the unweighted model (Table C), the points fell close to a straight line (through the origin) with slope 0.8. Thus, the predicted probabilities from the weighted model are systematically smaller, but not in a way that would have a substantial effect on the ordering of establishments according to their predicted probability.

**Table E. Final Weighted Logistic Regression Model, Fitted to the Augmented Data with Modified Adjusted Sampling Weights**

<b>Variable</b>	<b>Coefficient</b>	<b>Standard Error</b>	<b>P-Value</b>
Intercept	− 5.627	1.360	<.0001
Indicator_GT200	1.236	0.319	.0001
MinWhite_TenureRatio	− 1.068	0.493	.0303
FemMale_Diffi3	− 10.259	3.903	.0086
CompFemMale_TenureRatio	2.807	1.388	.0431

We recommend using the results from the unweighted model, for three main reasons. First, the main sample oversampled large establishments solely to ensure that it would contain a reasonable number of establishments with SD, and not be dominated by smaller establishments with no SD. Second, the EO Survey is not used to produce point estimates. And third, one of the variables in the model accounts for size (in two categories).